The Mediating Role of Explainable AI on Phishing Susceptibility

Masialeti Masialeti
Robert Morris University
Moon Township, USA
mxmst255@mail.rmu.edu

Wenli Wang
Dept. of Computer Information Systems
Robert Morris University
Moon Township, USA
wangw@rmu.edu

***Abstract*—Artificial intelligence (AI) is a double-edged sword for cybersecurity. Both cybercriminals and cybersecurity firms have taken advantages of Large Language Models (LLMs) to generate and defend attacks respectively. Although research has shown AI's impact on enhancing cybersecurity, there is limited research on the role of explainable AI (XAI) in cybersecurity. This research proposes the mediating role of XAI in AI-SETA (security education and training awareness) and phishing susceptibility. A measurement model is also proposed and tested empirically with 132 email users in the mining industry. Results show significant effects of AI-SETA and XAI on the reduction of phishing susceptibility.**

Keywords—artificial intelligence, phishing susceptibility, explainable AI, XAI, SETA, cybersecurity

I. INTRODUCTION

Cybercrime costs from data loss, financial theft, lower productivity, and Intellectual property (IP) infringement will reach $10.5 trillion USD by 2025 [1]. Email phishing is one of the major attack methods used by cybercriminals. Email phishing is a cyberattack in which attackers send emails with malicious links to entice users to click on them and compromise computer security [2]. Many companies invest in security education and training awareness (SETA) solutions to educate employees about cybersecurity attacks, especially email phishing. Cybersecurity trends, in both offense and defense, are transitioning toward artificial intelligence (AI) due to its capabilities and effects [3]. AI has become a double-edged sword in cybersecurity. Notably, cybercriminals have applied AI to make phishing emails look more legitimate. To counter, cybersecurity firms such as KnowBe4 are offering AI-enabled SETA to help companies raise employees' awareness in email phishing.

Since emails are mostly textual, the most applied AI technology in email phishing and its detection is Large Language Models (LLMs). LLMs are models that have natural language processing capabilities. Foundational LLMs with billions of parameters are typically built and offered by major technology companies. They typically pre-trained on extremely large datasets and provide responses to general users' prompts [4]. Cybersecurity companies like KnowBe4 have taken advantages of pre-trained foundational LLMs for AI-SETA in email phishing. For example, KnowBe4 fine-tunes pre-trained general LLMs with its own proprietary datasets on email phishing to enhance awareness training and offer phishing defense techniques for specific use cases.

LLMs have their limitations. For instance, hallucination is one of the major concerns of researchers, practitioners, and users regarding LLMs' applications in detecting email phishing. For most of the general users, LLMs' algorithms are "black boxes"; they don't know how AI algorithms work and how these algorithms can detect phishing emails. Hence, explainable artificial intelligence (XAI), which aims to help humans comprehend how AI algorithms work, becomes important to help build user trust in AI. For instance, XAI in AI-SETA for phishing detection can assist email users to understand how AI algorithms to classify phishing vs. legitimate emails.

XAI was described as a way to ensure that AI systems make accurate predictions and provide clear explanations for their predictions as sound decision support [5]. XAI can be referred to as tools, techniques and processes applied to AI model building and deployment to help make AI decision-makings transparent and understandable to human users [5]. From a behavioral perspective, XAI can also go beyond purely technical explanations but to focus on how human users perceive, interact with, and make sense of the AI systems and their decisions and outputs. In short, XAI can also be a measure in how well an AI user understands AI's final decisions.

Similarly, phishing susceptibility research also examines its related issues from both technical and behavioral perspectives. Phishing susceptibility, also known as "phishing vulnerability," is "the likelihood of an individual falling prey to a phisher, in which case s/he considers a phishing email to be legitimate, responds according to the email, and exposes private information, such as usernames, passwords, and credit card details" [19]. Many studies have been conducted to understand the causes of phishing [2][6][7][8] from both technical and behavioral perspectives, but limited research exists on AI applications in phishing detection. Since AI-SETA is a relatively new technique for phishing detection and XAI is also a relatively new concept, there is a lacking of research of the impact of AI-SETA and XAI on phishing susceptibility.

The purpose of this research is to understand the impact of AI-SETA and XAI on phishing susceptibility. It addresses the gap in the literature by investigating the interrelationships among AI-SETA, explainable AI (XAI), and phishing susceptibility, and proposing research and measurement models to test these interrelationships.

The remainder of this paper is organized as follows: Section II presents a literature review of the constructs and theories. Section III proposes the research model and its hypotheses. Section IV proposes the measurement model and discusses the results of a quantitative study. Section V concludes with summary, implications, limitations, and future research directions.

II. Literature Review

A. AI-SETA

AI-SETA is the application of AI in security education and training awareness (SETA) solutions to educate technology users about cybersecurity attacks, including email phishing. AI is defined as a set of technologies, machines, or systems capable of emulating human performance, typically by learning to understand complex data that normally requires cognition [9]. AI's abilities have contributed to the rise of its use in cybersecurity [10]. For instance, AI can continuously learn about cybersecurity threats and risks using massive data. It can find relationships among threats such as "malicious files, suspicious IP addresses, or insiders" in seconds. It not only delivers unmatchable speed by humans but also eliminates time-consuming tasks. Many studies have shown that AI has effectively reduced the impact of cyberattacks [10][11]. Studies also show increased use of AI in SETA [12][13][14].

AI for phishing detection is relatively new and is mainly LLMs-based. General foundational LLMs, such as GPT, Gemini, Llama, and Claude, are fine-tuned to go beyond the general natural language processing work to serve specific use cases such as identifying legitimate vs. phishing emails. Many cybersecurity providers have adopted LLMs in their detection and defense products.

LLMs can be trained through various learning methods. They are typically trained on an extensive corpus of unlabeled data through self-supervised learning, like predicting the next word. Foundational LLMs with billions of parameters are typically built and offered to the market by major technology companies such as OpenAI and Google, etc. A specialized information technology company (e.g., a cybersecurity firm) can then fine-tune its chosen foundational model(s) with its own specialty datasets (e.g., large amount of phishing emails across companies and industries) for specific tasks (e.g., phishing detection). Such fine-tuning can be done through supervised learning on a collection of labeled datas from a particular downstream domain (e.g., public SST-2, MNLI, and QQP inside the GLUE benchmark in the earlier days of LLMs development) or proprietary datasets (e.g., emails labelled as legitimate or phishing). The fine-tuning can be a continuous process: a cybersecurity firm can continuously train and fine-tune its specialized LLMs with its clients' contextual data (e.g., language patterns, work context, etc.).

B. Explainable Artificial Intelligence (XAI)

Explainable artificial intelligence (XAI) can be referred to the design tools and techniques offered by an AI system that facilitate the transparency of the otherwise "black box" approach of AI algorithms. Here, explainability refers to "the details and reasons a model gives to make its functioning clear or easy to understand" [15]. XAI can also be a perception of an AI system that indicates the degree to which an individual can comprehend the ultimate choice made by AI [16][17]. Here, explainability refers more to the "understandability" of an AI system. Understandability can involve both sides of the model and the user. By way of illustration, comprehensibility and transparency are associated with the capacity of a model to be intrinsically understandable, whereas interpretability is a measurement of how well a human being can comprehend the decision that a model makes. This two-sided perspective on words used in the field of XAI highlights the significance of taking into consideration the cognitive abilities and objectives of the user in addition to the qualities of the model. In light of this, the user perspective is emphasized in this research as an essential component of XAI.

C. Phishing Susceptibility

Phishing was defined as "the practice of sending deceptive electronic communications to acquire private information from victims, results in significant financial losses to individuals and businesses" [18]. Phishing susceptibility, also known as "phishing vulnerability," is "the likelihood of an individual falling prey to a phisher, in which case s/he considers a phishing email to be legitimate, responds according to the email, and exposes private information, such as usernames, passwords, and credit card details" [19].

Phishing susceptibility has been a subject of study in behavioral information security research for many years [18]. Although studies showed that IT security knowledge is needed to prevent phishing, employees still frequently fall victim to phishing attempts [19]. Phishing susceptibility research is typically classified in three main categories: 1) user characteristics (e.g., psychological, behavioral, or demographics factors); 2) characteristics of a phishing message; and 3) interventions such as trainings or warnings. User and message characteristics are typically theorized as having a direct association with an individual's phishing susceptibility, while interventions are often theorized as moderating those relationships [19].

Several attributes make people susceptible to phishing. These include phishing victims' traits and attack features [20][21][22]. Numerous personality traits were found to affect information processing, which affects phishing risk [23]. How a person fits within the organization's information flows and how one's workgroup duties affects one's thinking – they both affect phishing susceptibility [19]. IT security awareness affects phishing susceptibility in addition to contextualized, multi-level information processing [18]. Lack of technical skill in spotting phishing makes people vulnerable [8]. Internet anxiety increases phishing susceptibility [18]. When the link was textual but the source unclear, lack of focus had a substantial effect on phishing susceptibility [18]. When evaluated independently, risk propensity had a significant main effect in the known source and text link condition [18]. Participating in online communities minimizes phishing risk [18]. Unexpected email links with numbers are far less likely to be clicked [18]. Liking the sender increases receptivity [22]. No significant age or gender effect was found on phishing vulnerability [18].

Hence, the literature shows that much research exists with regard to phishing victims' traits and attack features. As AI-SETA for phishing detection is a relatively new technique and XAI is a relatively new construct for cybersecurity research, their relationships to phishing susceptibility need to be studied. This research aims to examine how AI-SETA training on phishing awareness reduces user phishing susceptibility behavior and what the role of XAI plays in the relationship.

III. Research Model

Fig. 1 illustrates the research model on AI-SETA and phishing susceptibility where XAI serves as a mediator. Research has shown that XAI enhances trust in AI-assistant decision making [24]. Trust often serves as a mediator, such as for the relationship between e-commerce website

determinants and online shopping behaviors [25] and for the relationship between satisfaction and continuous knowledge sharing intention [26]. Hence, XAI is proposed to play the mediating role for the AI-SETA's effect on reducing phishing susceptibility.

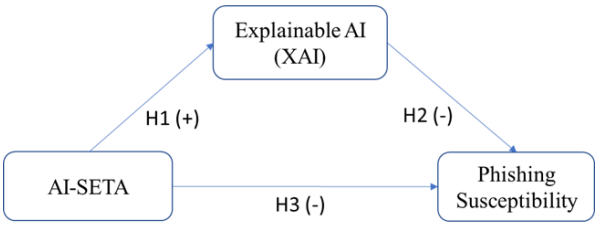

Fig. 1. Research model on AI-SETA and phishing susceptibility, where explainable AI serves as a mediator.

Having XAI as a mediator is important because the "black-box" approach of an AI system may mislead a user to over trust the AI system. Over trust due to the absence of explanations might lead to erroneous decisions, which would in turn lead to users massively distrusting the AI system provider – and trust repair after damage is usually very difficult [24]. XAI demands for extra efforts from AI system providers and require more user education to enhance their understandability. These XAI actions seem to be a cost but would have long-term benefits that overcome the initial cost.

The following hypotheses are derived from the proposed research model:

H1: AI-SETA training has a positive effect on explainability of AI.

H2: Explainability of AI has a negative effect on phishing susceptibility.

H3: AI-SETA training has a negative effect on phishing susceptibility partially through AI-SETA's effect on explainability of AI.

IV. Measurement Model and Empirical Result

Table I shows the measurement model. All definitions and measurement items are based on the literature. Applying instruments from the existing literature helps ensure the internal validity and reliability of the measurements.

TABLE I. Measurement Model of AI-SETA, XAI and PS.

Constructs	Definition	Measurement Items	Scale	Literature
AI-SETA	The inter-related activities organizations employ to foster employees' understanding and compliance with information security policies and guidelines using artificial intelligence features [27][12].	1) I understand that AI SETA increases my knowledge and skills in Cybersecurity on how not to click on link in phishing email.	5-point Likert scale	[28]
		2) I understand that AI SETA increases my knowledge and skills in Cybersecurity on how not to submit credentials		
		3) I understand that AI SETA increases my knowledge and skills in Cybersecurity on how report suspected phish.		
Explainable AI (XAI)	XAI is a characteristic of an artificial intelligence system that indicates the degree to which an individual can comprehend the ultimate choice made by AI [16][17].	1) I am confident in the AI feature to indicate whether an email is a phish.	5-point Likert scale	[29]
		2) The AI feature is accessible to me to check and see if an email is a phish.		
		3) The AI feature provides for privacy and confidentiality on the phishing and does not disclose the internal relationships of a trained AI model to unauthorized third parties.		
Phishing Susceptibility	The likelihood of an individual falling prey to a phisher, mistaking a phishing email to be legitimate, responding to the email, and exposing private information [19].	1) I am at risk of becoming victimized by phishing attacks.	5-point Likert scale	[30]
		2) It is likely that I will become victimized by phishing attacks.		
		3) It is possible that I will become victimized by phishing attacks.		
		4) My chances of getting phished are great.		

Note: All measurement items are in 5-point Likert-scale: 1 – strongly disagree; 2 – disagree; 3 – neither disagree or agree; 4 – Agree; 5 – strongly agree.

The research model and its measurement were tested through an empirical study of 132 of underground miners working for an international mining company in Spring 2025. The participant demographics is representative of the underground miner population.

Partial Least Square (PLS) Structural Equation Modeling (SEM) path modeling was used in SmartPLS 3.0 to evaluate data and test hypotheses. SEM is applied to examine the measurement model and analyze the structural model.

The measurement model showed high convergent validity and reliability. Confirmatory factor analysis (CFA) also assessed measure reliability and validity, concentrating on convergent and discriminant validity.

Hypothesis testing showed that all hypotheses were supported. There is significant relationship between AI-SETA and XAI. XAI significantly reduces phishing susceptibility. And AI-SETA also reduces phishing susceptibility through the partial mediating effect of XAI.

V. Conclusion, Implications, and Future Work

We proposed a research model on examining the interrelationships among AI-based security education and training awareness, explainable AI, and phishing susceptibility. We also designed a measurement model for the constructs. We used questionnaire survey data to confirm the validity and reliability of the measurements and to test the research model. AI-SETA's and explainable AI's negative effects on phishing susceptibility were both supported and statistically significant.

Hackers' use of phishing and cybersecurity crime's cost remain major challenges for cybersecurity practice and research [2]. Companies are using AI to reduce cybersecurity concerns like phishing [31]. AI's performance benefits continue to encourage its use in popular systems like email [32]. This research demonstrated the impacts of AI-SETA and explainable AI (XAI) on reducing phishing susceptibility. Specifically, our study extends the literature by providing the theories and empirical evidence on the role of explainable AI in reducing phishing susceptibility behavior. Explainable AI for phishing detection is not just hyped but as empirical evidence showed that there is statistical significance in the relationships. This study advances behavioral information security with explainable AI knowledge.

Our study also recommends organizations to choose AI systems that have enhanced explainability, especially in

cybersecurity applications. Email phishing detection AI should also explain how and why the AI systems can detect phishing. Such explanations help users to build trust in AI systems and decrease their susceptibility of clicking on harmful links.

The study has limitations. Our data size can be larger and the data can be more diverse such as including open-pit miners as well. The differentiation of underground vs. open-pit miners is important as underground miners are in higher risk environment and one miner's security weakness can potentially jeopardize many others' safety. Hence underground miners require even more heightened risk awareness. Email phishing susceptibility is an indicator of user's mindfulness and building mindfulness through AI-SETA can have an implicit impact on physical safety as well.

Future studies should explore the use of mixed method approach to gather triangulate findings of the quantitative study with qualitative data. Future study can also explore experimental study of the impact of enhanced explainability of AI on cybersecurity.

[1] eSentire Inc. "2023 official cybercrime report," eSentire. January 2024. [online]. Availale: https://www.esentire.com/resources/library/2023--official-cybercrime-report

[2] J. Wang, Li, Y., and Rao, H. R. "Overconfidence in phishing email detection," Journal of the Association for Information Systems, vol. 17, pp. 759–783. 2016. [online]. Available: https://doi.org/10.17705/1jais.00442

[3] T. Kim and Park, Y., "Artificial intelligence and firm performance," In Academy of Management Proceedings, pp. 14361. 2021.

[4] H. Zhao, Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. "Explainability for large language models: A survey," ACM Transactions on Intelligent Systems and Technology, vol. 15, pp. 1–38. 2024. [online]. Available: https://doi.org/10.1145/3639372

[5] W. Saeed and Omlin, C., "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," Knowledge-Based Systems, vol. 263. March 2023. [online]. Available: https://doi.org/10.1016/j.knosys.2023.110273.

[6] K. W. Hong, Kelley, C. M., Tembe, R., Murphy-Hill, E., and Mayhorn, C. B., "Keeping up with the Joneses: Assessing phishing susceptibility in an email task," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 57, pp. 1012-1016. September, 2013.

[7] P. Kumaraguru, Sheng, S., Acquisti, A., Cranor, L. F., and Hong, J. "Teaching Johnny not to fall for phish," ACM Transactions on Internet Technology, vol. 10, pp. 1-31. 2010.

[8] R. Dhamija, Tygar, J. D., and Hearst, M. "Why phishing works," Proceedings of the Conference on Human Factors in Computing Systems. 2006.

[9] R. E. Bawack, Wamba, S. F., and Carillo, K. "Artificial intelligence in practice: Implications for IS research," Proceeding of Americas Conference on Information Systems. 2019. [online]. Available: http://dblp.uni-trier.de/db/conf/amcis/amcis2019.html#BawackWC19

[10] IBM. "Artificial intelligence (AI) cybersecurity," 2023. [online]. Available: https://www.ibm.com/ai-cybersecurity

[11] O. Sharma, Sharma, A., and Kalia, A. "Windows and IoT malware visualization and classification with deep CNN and Xception CNN using Markov images," Journal of Intelligent Information Systems, vol. 60, pp. 349–375. 2022. [online]. Available: https://doi.org/10.1007/s10844-022-00734-4

[12] M. F. Ansari, Sharma, P. K., and Dash, B., "Prevention of phishing attacks using AI-Based cybersecurity awareness training," International Journal of Smart Sensors and Ad Hoc Networks, pp. 61–72. 2022. [online]. Available: https://doi.org/10.47893/ijssan.2022.1221

[13] T. E. Gasiba, Lechner, U., and Pinto-Albuquerque, M. "Sifu - a cybersecurity awareness platform with challenge assessment and intelligent coach," Cybersecurity, 3(1). 2020. https://doi.org/10.1186/s42400-020-00064-4

[14] M. M. Yamin, Shukla, A., Ullah, M., and Katt, B. "ADAPT-Automated Defence TrAining PLATForm in a Cyber range," In Lecture notes in networks and systems, pp. 184–203. 2023. [online]. Available: https://doi.org/10.1007/978-3-031-31153-6_17

[15] A. B. Arrieta, Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., and Benjamins, R., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115. 2020.

[16] A. Adadi and Berrada, M., "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160. 2018. [online]. Available: https://doi.org/10.1109/ACCESS.2018.2870052

[17] C. Collins, Dennehy, D., Conboy, K., and Mikalef, P. "Artificial intelligence in information systems research: A systematic literature review and research agenda," International Journal of Information Management, vol. 60. 2021. [online]. Available: https://doi.org/10.1016/j.ijinfomgt.2021.102383

[18] G. D. Moody, Galletta, D. F., and Dunn, B. K., "Which phish get caught? An exploratory study of individuals' susceptibility to phishing," European Journal of Information Systems, vol. 26, pp. 564-584. 2017. [online]. Available: https://doi.org/10.1057/s41303-017-0058-x

[19] R. Wright, Johnson, S. L., and Kitchens, B. "Phishing susceptibility in context: A multilevel information processing perspective on deception detection," Management Information Systems Quarterly, vol. 47, pp. 803–832. 2023. [online]. Available: https://doi.org/10.25300/misq/2022/16625

[20] R. Wright, and Marett, K. "The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived," Journal of Management Information Systems, vol. 27, pp. 273–303. 2010. [online]. Available: https://doi.org/10.2753/mis0742-1222270111

[21] R. Wright, Chakraborty, S., Basoglu, A., amd Marett, K. "Where did they go right? Understanding the deception in phishing communications," Group Decision and Negotiation, vol. 19, pp. 391–416. 2010. [online]. Available: https://doi.org/10.1007/s10726-009-9167-9

[22] R. Wright, Jensen, M. L., Thatcher, J. B., Dinger, M., and Marett, K., "Research note - influence techniques in phishing attacks: An examination of vulnerability and resistance," Information Systems Research, vol. 25, pp. 385–400. 2014. [online]. Available: https://doi.org/10.1287/isre.2014.0522

[23] E. D. Frauenstein and Flowerday, S. "Susceptibility to phishing on social network sites: A personality information processing model," Computers & Security, vol. 94. (2020). [online]. Available: https://doi.org/10.1016/j.cose.2020.101862

[24] B. Leichtmann, Humer, C., Hinterreiter, A., Streit, M., and Mara, M., "Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task," Computers in Human Behavior, vol. 139. 2023. [online]. Available: https://doi.org/10.1016/j.chb.2022.107539

[25] S. Ata, Baydas, A., and Yasar, M. E., "The relationships between determinants of shopping sites and customer e-trust, purchase intention, satisfaction, and repurchase," Journal of Economics and Administrative Sciences, vol. 22, pp. 324-349. 2021. [online]. Avalaible: https://doi.org/10.37880/cumuiibf.979417

[26] A. F. Hashima, and Tanb, F. B., "The mediating role of trust and commitment on members' continuous nowledge sharing intention : A commitment-trust theory perspective," Business, Psychology, Computer Science, 2014.

[27] W. Yaokumah, Walker, D. O., and Kumah, P., "SETA and security behavior," Journal of Global Information Management, vol. 27, pp. 102–121. 2019. [online]. Available: https://doi.org/10.4018/jgim.2019040106

[28] C. W. Yoo, Sanders, G. L., and Cerveny, R. P., "Exploring the influence of flow and psychological ownership on security education, training and awareness effectiveness and security compliance," Decision Support Systems, vol. 108, pp. 107–118. 2018. [online]. Available: https://doi.org/10.1016/j.dss.2018.02.009

[29] M. Masialeti, Talaei-Khoei, A., and Yang, A. "Revealing the role of explainable AI: How does updating AI applications generate agility-driven performance?" International Journal of Information Management, vol. 77, 2024. [online]. Available: https://doi.org/10.1016/j.ijinfomgt.2024.102779

[30] R. Chen, Gaia, J., and Rao, H. R., “An examination of the effect of recent phishing encounters on phishing susceptibility,” Decision Support Systems, vol. 133. 2020. [online]. Available: https://doi.org/10.1016/j.dss.2020.113287

[31] I. H. Sarker, “Machine learning: algorithms, Real-World applications and research directions,” SN Computer Science, vol. 3. 2021. [online]. Available: https://doi.org/10.1007/s42979-021-00592-x

[32] R. K. Behera, Bala, P. K., and Rana, N. P. “Creation of sustainable growth with explainable artificial intelligence: An empirical insight from consumer-packaged goods retailers,” Journal of Cleaner Production, vol. 399. 2023. [online]. Available: https://doi.org/10.1016/j.jclepro.2023.136605